EFFICIENT DATA UTILITY FROM ANONYMIZED PUBLISHED DATA USING GRAPHICAL REPRESENTATION

Sasikala.G[1], Hemamalini.J[2]
[1]Assistant Professor, [2]Research Scholar
Department of Computer Science, Adhiparaasakthi College of Arts and Science
Kalavai, TamilNadu, India

ABSTRACT: In recent years privacy preservation micro data publishing has gained wide popularity. Numerous anonymization techniques are used, namely generalization, bucketization and slicing are designed for privacy preserving micro data publishing. In generalization loses huge amount of information, peculiarly for high-dimensional data. In Bucketization, does not protect membership revealing and does not support for data clear separation between quasi-identifying attributes and sensitive attributes. In slicing, this partitions the data both horizontally and vertically. We show that slicing preserves better data utility than generalization and can be used for membership revealing protection. But it remains an open problem on how to use the anonymized data. To solve this issue Graphical representation method is used to efficient data utilization and data mining is approached.

Keywords: Privacy preservation, data anonymization, data publishing, data Utilization.

I. INTRODUCTION

PRIVACY-PRESERVING Publishing of Micro data

Micro data each records contains information about individual information, such as a person, a household, or an organization. Numerous micro data anonymization techniques have been proposed. The most notorious ones are generalization, for k-anonymity and bucketization, for diversity. In both approaches, attributes are distributed into three categories:

- Some attributes are identifiers that can uniquely identify an individual, such as Name or Social Security Number;
- Some attributes are Quasi Identifiers (QI), which the adversary may already know (possibly from other publicly available databases) and which, when taken together, can potentially identify an individual, e.g., Birth date, Sex, and Zip code, Route ,Disease, Animal reservoir;
- Some attributes are Sensitive Attributes (SAs), which are unknown to the adversary and are considered sensitive, such as Disease and Salary. The two techniques differ in the next step. Generalization transforms the QI-values in each bucket into "less specific but semantically consistent" values so that tuples in the same bucket cannot be distinguished by their QI values.

MOTIVATION OF SLICING - K-ANONYMITY

First, generalization for k-anonymity suffers from the curse of dimensionality. In order for generalization to be effective, records in the same bucket must be close to each other so that generalizing the records would not lose too much information. However, in high dimensional data, most data points have similar distances with each other, forcing a great amount of generalization to satisfy k-anonymity even for relatively small k's. Second, in order to perform data analysis or data mining tasks on the generalized table, the data analyst has to make the uniform distribution assumption that every value in a generalized interval/set is equally possible, as no other distribution assumption can be justified. his significantly reduces the data utility of the generalized data. Third, because each attribute is generalized separately, correlations between different attributes are lost. In order to study attribute correlations on the generalized table, the data analyst has to assume that every possible combination of attribute values is equally possible.

BUCKETIZATION

Bucketization has better data utility than generalization, it has several limitations. First, bucketization does not prevent membership disclosure. Because bucketization publishes the QI values in their original forms, an adversary can find out whether an individual has a record in the published data or not. A micro data (e.g., census data) usually contains many other attributes besides those three attributes. This means that the membership information of most individuals can be inferred from the bucketized table. Second, bucketization requires a clear separation between QIs and SAs. However, in many data sets, it is unclear which attributes are QIs and which are SAs. Third, by separating the sensitive attribute from the QI attributes, bucketization breaks the attribute correlations between the QIs and the SAs.

II. LITERATURE SURVEY

The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing many privacy protection algorithms rely on generalization and suppression of "quasiidentifier" attributes such as ZIP code and birthdate. Their objective is usually syntactic sanitization: for example, k-anonymity requires that each "quasi-identifier" tuple appear in at least k records, while `-diversity requires that the distribution of sensitive attributes for each quasi-identifier have high entropy. The utility of sanitized data is also measured syntactically, by the number of generalization steps applied or the number of records with the same quasi-identifier.

On the Anonymization of Sparse High-Dimensional Data The objective is to enforce privacy-preserving paradigms, such as k-anonymity and -diversity, while minimizing the information loss incurred in the anonymizing process (i.e. maximize data utility). However, existing techniques adopt an indexing- or clustering based approach, and work well for fixed-schema data, with low dimensionality. Nevertheless, certain applications require privacy-preserving publishing of transaction data (or basket data), which involves hundreds or even thousands of dimensions, rendering existing methods unusable. We employ a particular representation that captures the correlation in the underlying data, and facilitates the formation of anonymized groups with low information loss. We propose an efficient anonymization algorithm based on this representation.

Modeling and Integrating Background Knowledge in Data Anonymization The importance of considering the adversary’s background knowledge when reasoning about privacy in data publishing. However, it is very difficult for the data publisher to know exactly the adversary’s background knowledge. Existing work cannot satisfactorily model background knowledge and reason about privacy in the presence of such knowledge. This paper presents a general framework for modeling the adversary’s background knowledge using kernel estimation methods. This framework subsumes different types of knowledge (e.g., negative association rules) that can be mined from the data. Under this framework, we reason about privacy using Bayesian inference techniques and propose the skyline (B, t)- privacy model, which allows the data publisher to enforce privacy requirements to protect the data against adversaries with different levels of background knowledge.

Using Anonym zed Data for Classification In recent years, anonymization methods have emerged as an important tool to preserve individual privacy when releasing privacy sensitive data sets. This interest in anonymization techniques has resulted in a plethora of methods for anonymizing data under different privacy and utility assumptions. At the same time, there has been little research addressing how to effectively use the anonymized data for data mining in general and for distributed data mining in particular. A new approach for building classifiers using anonymized data by modeling anonymized data as uncertain data. we do not assume any probability distribution over the data. Instead, we propose collecting all necessary statistics during anonymization and releasing these together with the anonymized data. We show that releasing such statistics does not violate anonymity. Experiments spanning various alternatives both in local and distributed data mining settings reveal that our method performs better than heuristic approaches for handling anonymized data.

T-Closeness: Privacy Beyond k-Anonymity and l-Diversity The k-anonymity privacy requirement for publishing microdata requires that each equivalence class (i.e., a set of records that are indistinguishable from each other with respect to certain "identifying" attributes) contains at least k records. Recently, several authors have recognized that k-anonymity cannot prevent attribute disclosure. The notion of l-diversity has been proposed to address this; l-diversity requires that each equivalence class has at least l well-represented values for each sensitive attribute. We propose a novel privacy notion called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). We choose to use the earth mover distance measure for our t-closeness requirement. We discuss the rationale for t-closeness and illustrate its advantages through examples and experiments.

Injector: Mining Background Knowledge for Data Anonymization existing work on privacy-preserving data publishing cannot satisfactorily prevent an adversary with background knowledge from learning important sensitive information. The main challenge lies in modeling the adversary's background knowledge. We propose a novel approach to deal with such attacks. In this approach, one first mines knowledge from the data to be released and then uses the mining results as the background knowledge when anonymizing the data. The rationale of our approach is that if certain facts or background knowledge exist, they should manifest themselves in the data and we should be able to find them using data mining techniques. One intriguing aspect of our approach is that one can argue that it improves both privacy and utility at the same time, as it both protects against background knowledge attacks and better preserves the features in the data.

III. DESIGN AND IMPLEMENTATION

The basic idea of slicing is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization. Slicing preserves utility because it groups highly correlated attributes together, and preserves the correlations between such attributes. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are infrequent and thus identifying. Note that when the data set contains QIs and one SA, bucketization has to break their correlation; slicing, on the other hand, can group some QI attributes with the SA, preserving attribute correlations with the sensitive attribute. Slicing first partitions attributes into columns. Each column contains a subset of attributes. This vertically partitions the table. Slicing also partition tuples into buckets. Each bucket contains a subset of tuples. This horizontally partitions the table. Within each bucket, values in each column are randomly permutated to break the linking between different columns.

Slicing preserves more information than such a local recoding approach, assuming that the same tuple partition is used. We achieve this by showing that slicing is better than the following enhancement of the local recoding approach. Rather than using a generalized value to replace more specific attribute values, one uses the multiset of exact values in each bucket. In slicing, one group correlated attributes together in one column and preserves their correlation.

Another important advantage of slicing is its ability to handle high-dimensional data. By partitioning attributes into columns, slicing reduces the dimensionality of the data. Each column of the table can be viewed as a sub table with a lower dimensionality. The idea of slicing is to achieve a better trade-off between privacy and utility by preserving correlations between highly correlated attributes and breaking correlations between uncorrelated attributes.

Privacy Threats

When publishing microdata, there are three types of privacy disclosure threats.

- Membership disclosure
- identity disclosure
- attribute disclosure

MEMBERSHIP DISCLOSURE

When the data set to be published is selected from a large population and the selection criteria are sensitive (e.g., only diabetes patients are selected), one needs to prevent adversaries from learning whether one’s record is included in the published data set.

IDENTITY DISCLOSURE

Identity disclosure occurs when an individual is linked to a particular record in the released table. In some situations, one wants to protect against identity disclosure when the adversary is uncertain of membership. In this case, protection against membership disclosure helps protect against identity disclosure. In other situations, some adversary may already know that an individual’s record is in the published data set, in which case, membership disclosure protection either does not apply or is insufficient.

ATTRIBUTE DISCLOSURE

Attribute disclosure, occurs when new information about some individuals is revealed, i.e., the released data make it possible to infer the attributes of an individual more accurately than it would be possible before the release. Similar to the case of identity disclosure, we need to consider adversaries who already know the membership information. Identity disclosure leads to attribute disclosure. Once there is identity disclosure, an individual is reidentified and the corresponding sensitive value is revealed. Attribute disclosure can occur with or without identity disclosure, e.g., when the sensitive values of all matching tuples are the same.

SLICING ALGORITHMS

- Attribute Partitioning
- Column Generalization
- Tuple Partitioning

Attribute Partitioning

This algorithm partitions attributes so that highly correlated attributes are in the same column. This is good for both utility and privacy. In terms of data utility, grouping highly correlated attributes preserves the correlations among those attributes. In terms of privacy, the association of uncorrelated attributes presents higher identification risks than the association of highly correlated attributes because the association of uncorrelated attributes values is much less frequent and thus more identifiable. Therefore, it is better to break the associations between uncorrelated attributes, in order to protect privacy.

Column Generalization

Tuples are generalized to satisfy some minimal frequency requirement. Bucketization provides the same level of privacy protection as generalization, with respect to attribute disclosure. Although column generalization is not a required phase, it can be useful in several aspects.

- Column generalization may be required for identity/membership disclosure protection. If a column value is unique in a column (i.e., the column value appears only once in the column), a tuple with this unique column value can only have one matching bucket.
- When column generalization is applied, to achieve the same level of privacy against attribute disclosure, bucket sizes can be smaller. While column generalization may result in information loss, smaller bucket-sizes allow better data utility.

Tuple Partitioning The main part of the tuple-partition algorithm is to check whether a sliced table satisfies L-diversity. For each tuple t, the algorithm maintains a list of statistics L[t] about t’s matching buckets. Each element in the list L[t] contains statistics about one matching bucket B: the matching probability p(t,B) and the distribution of candidate sensitive values D(t,B).

MODULES DESCRIPTION

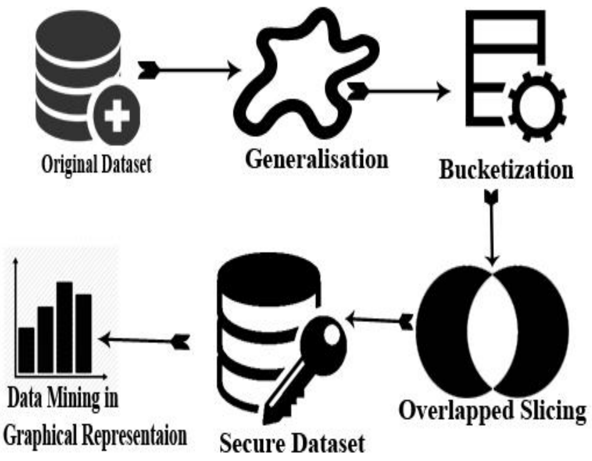


Figure 3.1 - Architecture Diagram

DATASET EXTRACTION: The dataset extraction module can be used to extract the dataset and it will be stored in the database for future use. Initially the dataset was selected, after that it will be split separate data and it can be stored in the table to the user database.

GENERALIZATION: Generalization module performs 2-anonymity process. In generalization approach we use the identifiers data and Quasi Identifiers. Here the attribute age is Identifiers, and gender is Quasi Identifiers. The generalization data can be retrieved from an original data. The dataset data’s are stored into two buckets.

BUCKETIZATION: Bucketization module can be performs 2-diversity process. In generalization approach we use the Quasi Identifiers. Here the attribute work class is attribute. The bucketization data can be retrieved from an original data. The dataset data's are stored into two buckets.

MULTI-SET GENERALIZATION: Multi-set generalization module performs 2-anonymity process. In multi-set generalization approach we use the identifiers data and Quasi Identifiers. Here the attribute age is Identifiers, and gender, work class are Quasi Identifiers. The multi-set generalization data can be retrieved from an original data. The dataset data's are stored into two buckets.

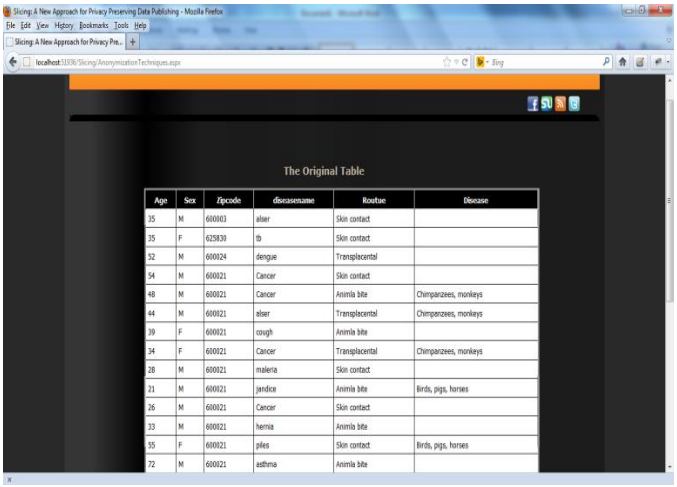
SLICING: Slicing partitions the data set both vertically and horizontally. Slicing preserves better data utility than generalization and can be used for membership disclosure protection. Here we using the following sub modules,

- Attribute partition and Columns
- Tuple Partition and Buckets
- Slicing
- Column Generalization
- Matching Buckets

GRAPH GENERATION:

Graph generation module can be used to find the classification accuracy between Original data, Generalization, Bucketization and Slicing. Slicing shows better accuracy than generalization. When the target attribute is the sensitive attribute, slicing even performs better than bucketization.

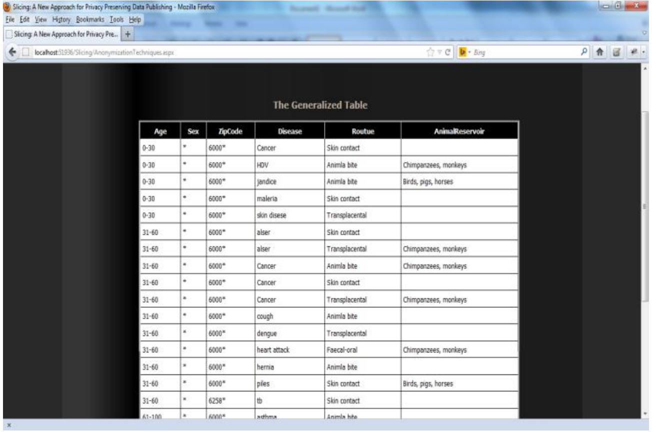
IV. RESULT


Slicing: A New Approach for Privacy Preserving Data Publishing - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Slicing: A New Approach for Privacy Pre...
localhost:51936/Slicing/AnonymizationTechniques.aspx
Bing

The Original Table

Age	Sex	Zipcode	diseasename	Routue	Disease
35	M	600003	alser	Skin contact	
35	F	625830	tb	Skin contact	
52	M	600024	dengue	Transplacental	
54	M	600021	Cancer	Skin contact	
48	M	600021	Cancer	Animla bite	Chimpanzees, monkeys
44	M	600021	alser	Transplacental	Chimpanzees, monkeys
39	F	600021	cough	Animla bite	
34	F	600021	Cancer	Transplacental	Chimpanzees, monkeys
28	M	600021	maleria	Skin contact	
21	M	600021	jandice	Animla bite	Birds, pigs, horses
26	M	600021	Cancer	Skin contact	
33	M	600021	hernia	Animla bite	
55	F	600021	piles	Skin contact	Birds, pigs, horses
72	M	600021	asthma	Animla bite	

Figure 4.1- Original Data


Slicing: A New Approach for Privacy Preserving Data Publishing - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Slicing: A New Approach for Privacy Pre...
localhost:51936/Slicing/AnonymizationTechniques.aspx
Bing

The Generalized Table

Age	Sex	ZipCode	Disease	Routue	AnimalReservoir
0-30	*	6000*	Cancer	Skin contact	
0-30	*	6000*	HDV	Animla bite	Chimpanzees, monkeys
0-30	*	6000*	jandice	Animla bite	Birds, pigs, horses
0-30	*	6000*	maleria	Skin contact	
0-30	*	6000*	skin disese	Transplacental	
31-60	*	6000*	alser	Skin contact	
31-60	*	6000*	alser	Transplacental	Chimpanzees, monkeys
31-60	*	6000*	Cancer	Animla bite	Chimpanzees, monkeys
31-60	*	6000*	Cancer	Skin contact	
31-60	*	6000*	Cancer	Transplacental	Chimpanzees, monkeys
31-60	*	6000*	cough	Animla bite	
31-60	*	6000*	dengue	Transplacental	
31-60	*	6000*	heart attack	Faecal-oral	Chimpanzees, monkeys
31-60	*	6000*	hernia	Animla bite	
31-60	*	6000*	piles	Skin contact	Birds, pigs, horses
31-60	*	6258*	tb	Skin contact	
61-100	*	6000*	asthma	Animla bite	

Figure 4.2- Generalized Data

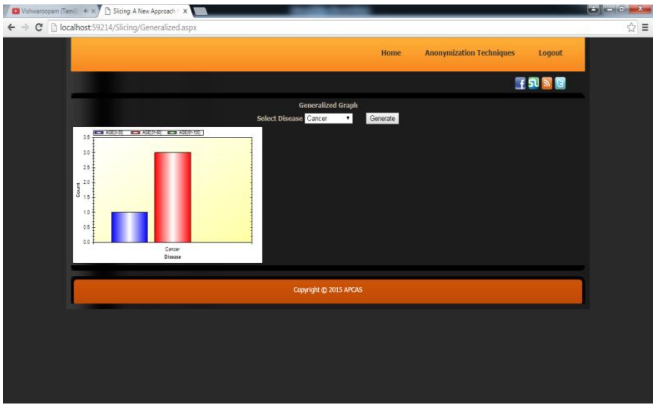
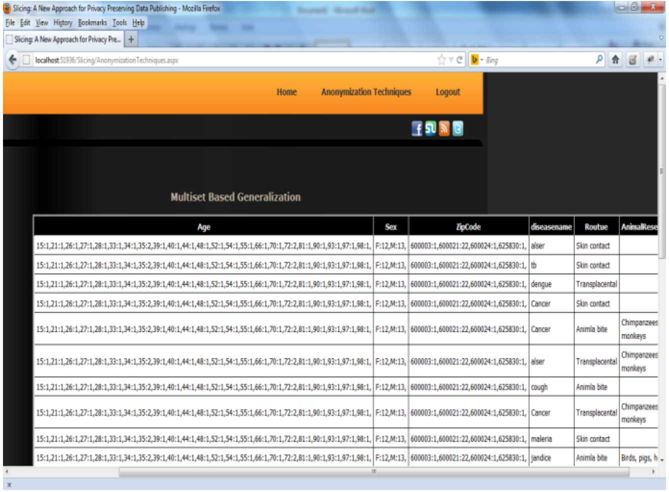


Figure 4.3- Generalized Graph


Slicing: A New Approach for Privacy Preserving Data Publishing - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Slicing: A New Approach for Privacy Pre...
localhost:51936/Slicing/AnonymizationTechniques.aspx
Bing
Home
Anonymization Techniques
Logout

Multiset Based Generalization

Age	Sex	ZipCode	diseasename	Routue	AnimalRese
15:1,21:1,26:1,27:1,28:1,33:1,34:1,35:2,39:1,40:1,44:1,48:1,52:1,54:1,55:1,66:1,70:1,72:2,81:1,90:1,93:1,97:1,98:1,	F:12,M:13,	600003:1,600021:22,600024:1,625830:1,	alser	Skin contact	
15:1,21:1,26:1,27:1,28:1,33:1,34:1,35:2,39:1,40:1,44:1,48:1,52:1,54:1,55:1,66:1,70:1,72:2,81:1,90:1,93:1,97:1,98:1,	F:12,M:13,	600003:1,600021:22,600024:1,625830:1,	tb	Skin contact	
15:1,21:1,26:1,27:1,28:1,33:1,34:1,35:2,39:1,40:1,44:1,48:1,52:1,54:1,55:1,66:1,70:1,72:2,81:1,90:1,93:1,97:1,98:1,	F:12,M:13,	600003:1,600021:22,600024:1,625830:1,	dengue	Transplacental	
15:1,21:1,26:1,27:1,28:1,33:1,34:1,35:2,39:1,40:1,44:1,48:1,52:1,54:1,55:1,66:1,70:1,72:2,81:1,90:1,93:1,97:1,98:1,	F:12,M:13,	600003:1,600021:22,600024:1,625830:1,	Cancer	Skin contact	
15:1,21:1,26:1,27:1,28:1,33:1,34:1,35:2,39:1,40:1,44:1,48:1,52:1,54:1,55:1,66:1,70:1,72:2,81:1,90:1,93:1,97:1,98:1,	F:12,M:13,	600003:1,600021:22,600024:1,625830:1,	Cancer	Animla bite	Chimpanzees monkeys
15:1,21:1,26:1,27:1,28:1,33:1,34:1,35:2,39:1,40:1,44:1,48:1,52:1,54:1,55:1,66:1,70:1,72:2,81:1,90:1,93:1,97:1,98:1,	F:12,M:13,	600003:1,600021:22,600024:1,625830:1,	alser	Transplacental	Chimpanzees monkeys
15:1,21:1,26:1,27:1,28:1,33:1,34:1,35:2,39:1,40:1,44:1,48:1,52:1,54:1,55:1,66:1,70:1,72:2,81:1,90:1,93:1,97:1,98:1,	F:12,M:13,	600003:1,600021:22,600024:1,625830:1,	cough	Animla bite	
15:1,21:1,26:1,27:1,28:1,33:1,34:1,35:2,39:1,40:1,44:1,48:1,52:1,54:1,55:1,66:1,70:1,72:2,81:1,90:1,93:1,97:1,98:1,	F:12,M:13,	600003:1,600021:22,600024:1,625830:1,	Cancer	Transplacental	Chimpanzees monkeys
15:1,21:1,26:1,27:1,28:1,33:1,34:1,35:2,39:1,40:1,44:1,48:1,52:1,54:1,55:1,66:1,70:1,72:2,81:1,90:1,93:1,97:1,98:1,	F:12,M:13,	600003:1,600021:22,600024:1,625830:1,	maleria	Skin contact	
15:1,21:1,26:1,27:1,28:1,33:1,34:1,35:2,39:1,40:1,44:1,48:1,52:1,54:1,55:1,66:1,70:1,72:2,81:1,90:1,93:1,97:1,98:1,	F:12,M:13,	600003:1,600021:22,600024:1,625830:1,	jandice	Animla bite	Birds, pigs, h

Figure 4.4- MultiSet Table

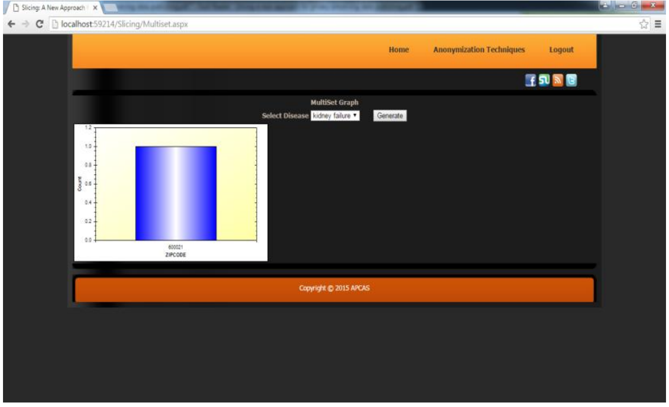
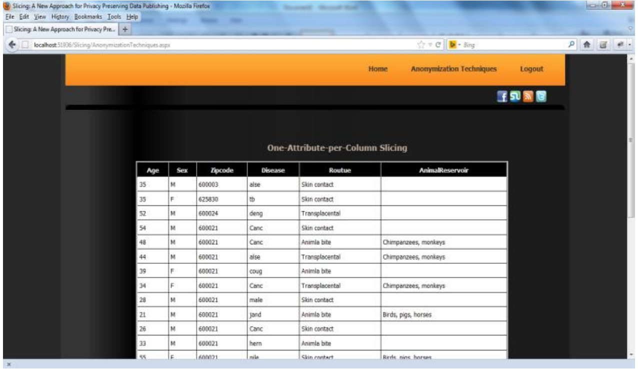


Figure 4.5- MultiSet Graph


Slicing: A New Approach for Privacy Preserving Data Publishing - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Slicing: A New Approach for Privacy Pre...
localhost:51936/Slicing/AnonymizationTechniques.aspx
Bing
Home
Anonymization Techniques
Logout

One-Attribute-per-Column Slicing

Age	Sex	Zipcode	Disease	Routue	AnimalReservoir
35	M	600003	alse	Skin contact	
35	F	625830	tb	Skin contact	
52	M	600024	deng	Transplacental	
54	M	600021	Canc	Skin contact	
48	M	600021	Canc	Animla bite	Chimpanzees, monkeys
44	M	600021	alse	Transplacental	Chimpanzees, monkeys
39	F	600021	coug	Animla bite	
34	F	600021	Canc	Transplacental	Chimpanzees, monkeys
28	M	600021	male	Skin contact	
21	M	600021	jand	Animla bite	Birds, pigs, horses
26	M	600021	Canc	Skin contact	
33	M	600021	hern	Animla bite	
55	F	600021	pile	Skin contact	Birds, pigs, horses

Figure 4.6- One attribute per column

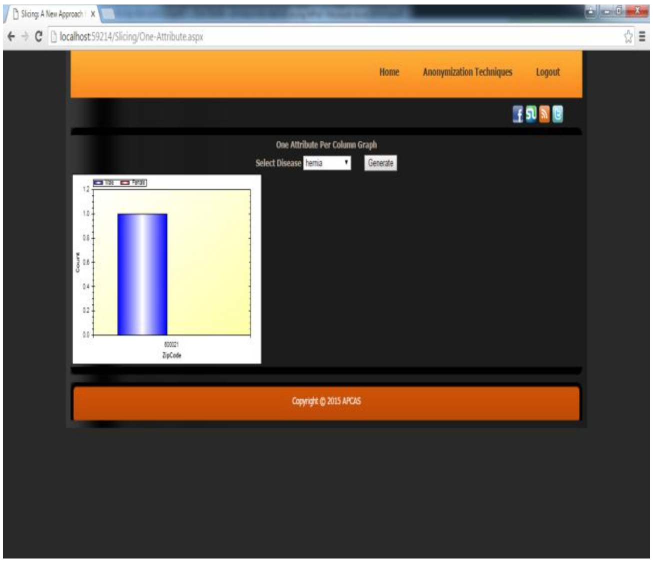
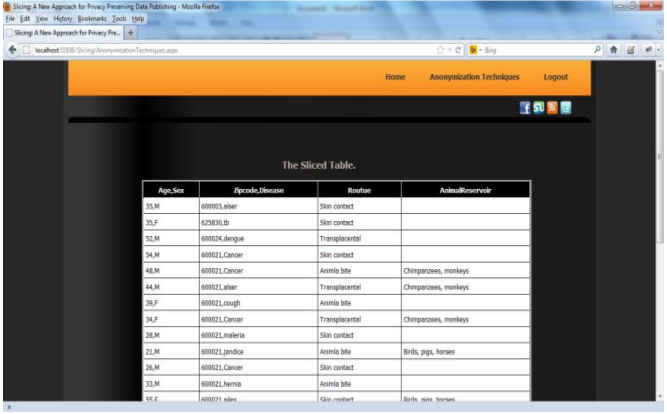


Figure 4.7- One attribute per column Graph


Slicing: A New Approach for Privacy Preserving Data Publishing - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Home
Anonymization Techniques
Logout

The Sliced Table.

Age,Sex	Zipcode,Disease	Routue	AnimalReservoir
35,M	600003,alser	Skin contact	
35,F	625830,tb	Skin contact	
52,M	600024,dengue	Transplacental	
54,M	600021,Cancer	Skin contact	
48,M	600021,Cancer	Animla bite	Chimpanzees, monkeys
44,M	600021,alser	Transplacental	Chimpanzees, monkeys
39,F	600021,cough	Animla bite	
34,F	600021,Cancer	Transplacental	Chimpanzees, monkeys
28,M	600021,maleria	Skin contact	
21,M	600021,jandice	Animla bite	Birds, pigs, horses
26,M	600021,Cancer	Skin contact	
33,M	600021,hernia	Animla bite	
55,F	600021,piles	Skin contact	Birds, pigs, horses

Figure 4.8- Sliced Table

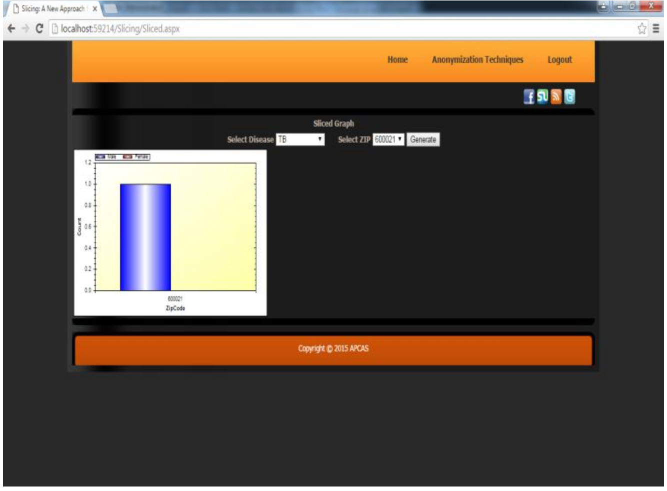


Figure 4.9- Sliced Graph

V. CONCLUSION

The limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. Slicing is used to prevent attribute disclosure and membership disclosure. Slicing preserves better data utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. The general methodology proposed by this work is that: before anonymizing the data, one can analyze the data characteristics and use these characteristics in data anonymization. The rationale is that one can design better data anonymization techniques when we know the data better.

Future Enhancement

Various number of anonymization techniques have been designed; it remains an open problem on how to use the anonymized data. In our experiments, we randomly generate the associations between column values of a bucket. This may lose data utility. Another direction is to design data mining tasks using the anonymized data computed by various anonymization techniques.

REFERENCE

[1] C. Aggarwal, “On k-Anonymity and the Curse of Dimensionality,” Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 901-909,2005.

[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim, “Practical Privacy: The SULQ Framework,” Proc. ACM Symp. Principles of Database Systems (PODS), pp. 128-138, 2005.

[3] J. Brickell and V. Shmatikov, “The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing,” Proc. ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008.

[4] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, “Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge,” Proc.Int’l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.

[5] H. Cramt’er, Mathematical Methods of Statistics. Princeton Univ. Press, 1948.

[6] I. Dinur and K. Nissim, “Revealing Information while Preserving Privacy,” Proc. ACM Symp. Principles of Database Systems (PODS),pp. 202-210, 2003

[7] C. Dwork, “Differential Privacy,” Proc. Int’l Colloquium Automata, Languages and Programming (ICALP), pp. 1-12, 2006.